Supplementary Material and Methods

Study system and sampling

The study was conducted in the natural populations of *Plantago lanceolata* in the Åland Island, SW of Finland. *Plantago lanceolata* is a globally occurring perennial herbaceous plant[1]. It is an obligate out-crosser with wind-dispersed pollen, and it is also capable of clonal reproduction via side rosettes [1]. In the Åland Islands, *P. lanceolata* typically grows on dry meadows, forming a highly fragmented network of approximately 4000 populations spanning an area of approximately 50 km by 70 km[2]. The size and location of the populations have been monitored annually since early 1990's as a part of the metapopulation studies of the Glanville fritillary butterfly *Melitaea cinxia*[2]. Due to the clonal reproduction and a seed bank, *P. lanceolata* populations rarely go extinct, and thus the spatial configuration of these populations is relatively stable[3].

We used a stratification process to select 20 focal populations from the database of the approximately 4000 natural populations of *P. lanceolata* for this study. First, we divided the populations into small and large ones based on the lower and upper quartiles of the population sizes (recorded as the area covered by the plant in square meters at each locality during annual surveys[2]). Second, based on the lower and upper quartiles of population connectivity measures (Table 1), we divided the populations to the ones with low connectivity and the others with high connectivity. Population connectivity is assumed to provide a powerful proxy for migration rates between populations[3]. After overlaying these populations onto the map of Åland, we randomly selected four populations with high or low size or connectivity, respectively, from five different areas of Åland.

We sampled the selected 20 populations in early June 2017 for their virus communities. We collected leaf samples from 20 randomly selected plants in each population, resulting in altogether 400 samples for small RNA sequencing. Each leaf sample consisted of three youngest leaves from the centre of the rosette, which were placed in 5 ml cryotubes and immediately frozen in liquid nitrogen, and subsequently stored in –80°C.

Processing the RNA samples and detecting virus communities

For identifying the viruses, RNA was extracted from the samples using phenol-chloroform-isoamylalcohol extraction, following a modified protocol of Chang et al.[4]. RNA yield and quality were measured using Nanodrop (Thermo Fisher Scientific Inc., Waltham, MA, USA). The small RNA (sRNA) sequencing-based virus detection was done based on Cuellar *et al*.[5]. sRNA (18-30 nucleotides) was sequenced using Illumina HiSeq (Illumina Inc., San Diego, CA, USA)]. The RNA samples were sequenced by a third-party company (FASTERIS https://www.fasteris.com/dna/) with insert size of 18 to 30 bases with an average library yield of 950 Mb. After quality check and adaptor removal, the obtained reads were *de novo* assembled to contigs and blastn and blastx searches were run against RVDBv10.2 database[6] using the VirusDetect pipeline[7] on all the samples separately. In BLAST searches we used default similarity 25% and 1e-5. All BLAST hits were manually checked to avoid false positives. We then assigned the VirusDetect hits to Operational Taxonomic Units (OTUs) by assigning all hits belonging to the same family to the same OTU. The virus OTUs were classified as plant infecting viruses, mycoviruses, or as other viruses according to their host range[8]. A presence-absence table per sample and per virus OTU was created based on the output results of the VirusDetect pipeline, henceforth referred to as the virus community matrix.

EXPLANATORY VARIABLES

Host-related variables

During sampling, we tagged the plants, recorded their locations with GPS (the centroid of all the sampled host populations was 60°11'57.3"N, 19°58'47.0"E), and recorded the size of the plants (number of leaves and flowers, and the length and width of the longest leaf on each plant, Table 1) to understand how plant size affects the occurrence of viruses. To understand how herbivory of the plant individual is linked with the occurrence of viruses, we also recorded signs of herbivore damage (suck- and bitemarks, holes, moth pupa, leaf miner or thrip damage and spittle bug eggs, all recorded separately as 0/1, see Table 1).

Habitat-related variables

As these viruses are not expected to be specialists of *P. lanceolata*[9], we wanted to test whether the diversity of the local plant communities affects virus community structure. Hence, we collected data on the vascular plant communities in each of the study populations between the 19th and 30th of June 2017. We established between three and six one-square-meter vegetation survey rectangles across each focal population, depending on its area. For each square, we counted the number of vascular plant species. We used the vegetation data to calculate the Shannon diversity index[10] for each population (Table 1).

Theory suggests pathogen emergence to increase at the interface of wild and cultivated areas[11] and previous studies have reported higher virus diversities near cultivated areas[12,13]. The proportion of agricultural area surrounding each study population was calculated from Corine Land Cover (CLC, version 2020_20u1) with QGIS[14] by creating a one-kilometre buffer zone around each study population following the patch borders and calculating the proportion of 20 m × 20 m pixels falling under agricultural land use category within this buffer zone (Table 1). To quantify the effect of host population size on the occurrence of viruses, we also estimated the coverage of *P. lanceolata* foliage in square meters in each population. Furthermore, to quantify the connectivity of a host population with respect to other populations, we calculated the Euclidian distances between populations, calibrated by the species dispersal capacity to measure the connectivity between host plant populations[3].

As many pathogens are known to be very sensitive to the surrounding climatic conditions[15], we assume that the preceding weather conditions affect the virus community. The weather observations for the host populations were obtained from the Finnish Meteorological Institut[16]. We calculated the number of severe winter days during the winter before the sampling season (i.e. number of days when temperature was < 0°C and with snow depth was < 5 cm in winter 2016-2017) and the sum over temperatures of the effective summer days during previous summer (2016) (Table 1).

Spatial variables

To account for spatial structure in the data, we included spatial variables implemented as Moran's eigenvector maps (MEMs). MEMs describe the spatial structure of data as a correlation among nearby locations in space, and they can readily be used as spatial predictors[17]. We calculated MEMs based on the GPS coordinates of the sampled host plants, with the assumption of positive autocorrelation. We estimated the significance of individual eigenvectors with a permutation test for the Moran's I statistic, which describes the level of autocorrelation in the data. We included all the eigenvectors with Moran's I >0.7, as well as the last significant one, thus representing both coarse and fine scale signal of significant spatial autocorrelation. We used the R environment[18] and packages 'spdep'[19] and 'adespatial'[20].

STATISTICAL METHODS

Descriptive analyses

To address our research question Q1, we began with descriptive analyses of the virus species co-occurrence structure. We looked into the nestedness of the data set visually by organising the virus community matrix based on overlap in virus presences among plants and decreasing fill[21]. We calculated the C-scores[22] for the virus community, at the level of individual host plants, as well as at the host plant population-level (Figure 2). The metacommunity of viruses is considered nested when the virus community in plants that host a few viruses represent a subset of the virus community in plants that host more viruses. We also described the structure of the virus communities visually within host populations as well as within individual plants for two most contrasting host populations in terms of their virus diversity and nestedness. We used the full dataset of 25 viruses (See Results) for these descriptive illustrations. To further characterize the co-occurrence structure of the virus communities, we calculated the numbers of unique pairwise virus co-occurrence combinations in host plants, as well as the unique pairs of virus species that never co-occur.

To address our research question Q1 regarding the relevant scales of virus diversity and co-occurrence, we looked into the relationship between the cumulative species richness and species co-occurrence patterns with respect to increasing area sampled by calculating the mean species–area–curve and mean species–coexistence–curves. The classic species–area–curve describes how the species richness accumulates with increasing area. We calculated a Type I curve[23], with a nested structure, with stepwise accumulation of species. The coexistence–area relationship describes the rate at which the number of coexisting species grows with increasing area, and reflects the scale-dependent operation of the mechanisms that contribute to coexistence[24]. Both curves were constructed by randomly selecting a starting point, i.e. a sampled host plant, and increasing the spatial scale and sample by always including the next closest plant to the species richness or co-occurrence calculation. We repeated this 100 times, every time with a different random starting point, and from these we calculated the mean curves. We also calculated the maximum amount of co-occurring pairs based on the number of species observed thus far, leading to a faster rise.

Markov Random Field networks

In order to address our study questions Q2-Q4 regarding the relative effects of host, habitat quality and weather characteristics and spatial structure of the host populations, and the direct and indirect associations between the viruses, we fitted both Unconditional (MRF) and Conditional Markov Random Field models (CRF) to the virus community data[25]. Markov Random Fields (MRFs) are graphical models, which can represent complex distributions as network graphs. These networks consist of *nodes* and *edges*, corresponding to the observed variables within the data, and to the probabilistic interactions between variables that need to be estimated.

The edge associations are undirected, meaning that the effect of one node on another is reciprocal. If there is no edge between two nodes in the estimated graph, these nodes are conditionally independent from one another, whereas if there is an edge, these nodes are conditionally dependent, *after* accounting for the other node effects in the graph model[26]. Here, the edge associations of the network describe the

direct pairwise associations between viruses in host plants. Conditional Random Fields allow for these dependencies among nodes to be further conditional on other covariates[25,26]. Hence, the values for the edge associations can change in the presence of these covariates, and the resulting graph model illustrates the pairwise associations between viruses in host plants, conditional not only on the rest of the virus community, but also on the covariates included in the model (Table 1).

The applied Markov Random Field modelling framework is described in detail by Clark *et al.*[25]. Briefly, we modelled the log-odds of detecting virus *i* given covariate *x* and occurrence of virus *j* with

$$\log\left(\frac{P(y_i = 1|y_{\setminus i}, x)}{1 - P(y_i = 1|y_{\setminus i}, x)}\right) = \alpha_{i0} + \beta_i^T x + \sum_{j:j\neq i} (\alpha_{ij0} + \beta_{ij}^T x)\, y_j \,,$$

where y_i is the vector of presences and absences of virus *i*; $y_{\setminus i}$ denotes the presences and absences of all other viruses except *i*; α_{i0} is the virus-level intercept; and β_i^T is the effect of covariate x on the occurrence probability of virus *i*. Parameters α_{ij0} and β_{ij}^T represent the associations between species, conditional on the occurrences of all the non-focal viruses (other than the focal virus *i*). Their interpretation is easiest explained by setting one parameter to zero: If, for example, $\alpha_{ij0} = 0$ but $\beta_{ij}^T \neq 0$, the occurrence probabilities of virus *i* and *j* are conditionally independent, after accounting for the effect of the covariate x, represented by β_{ij}^T, and the occurrences of other viruses. If on the other hand $\alpha_{ij0} \neq 0$ but $\beta_{ij}^T = 0$, the occurrence probabilities of virus *i* are conditionally dependent on species *j*, but this association does not vary with covariate x. Hence, if both $\alpha_{ij0} \neq 0$ and $\beta_{ij}^T \neq 0$, the occurrence probabilities of virus *i* and *j* are conditionally dependent, after accounting for the effect of the covariate x, represented by β_{ij}^T, and the occurrences of species *j*.

For fitting the MRF and CRF models, we used data on all viruses with at least 10 occurrences in the entire virus community matrix (i.e. minimum prevalence of 2.5% of sampling units). For understanding how different characteristics of the environment and the host affect the virus community, we included several explanatory variables in the model (Table 1), describing: 1) The level of s*patial autocorrelation* of the host populations (implemented as Moran's eigenvectors) 2) *habitat-related* characteristics,

namely the *quality* of the habitat of the host plants (the connectedness (*S*) of the focal *P. lanceolata* population to other populations, agricultural land use (percentage of the surrounding landscape) and the Shannon diversity of the local plant community, which have been demonstrated to influence virus occurrences in this system[27], and as the *weather* conditions of local populations (severity of the previous winter and temperature sum over the effective summer days during previous summer); as well as 4) *host-related* characteristics of the focal host plant individual (host population size, host plant individual size and signs of herbivory). See Table 1 for full details.

Altogether our dataset used for modelling consisted of 16 virus species and 16 explanatory variables (Table 1), resulting in 272 coefficients in each regression. To avoid overfitting, regularisation has been implemented in the method through least absolute shrinkage and selection operator (LASSO), forcing some regression coefficients to zero, and thus performing variable selection and reducing the risk of overfitting [25]. Regularisation is influenced by the scale of the covariates, and hence we scaled all our continuous variables to mean zero and standard deviation one. The CRF model is estimated with separate logistic regressions. To achieve an undirected network and symmetry within the coefficients of conditional dependence (so that $\alpha_{ij0} = \alpha_{ji0}$ and $\beta_{ij}^{T} = \beta_{ji}^{T}$) we take the mean of the corresponding estimates, which is the default setting of the applied algorithm[25].

We fitted six model variants in total: 1) an Unconditional Markov Random Field model (referred to as 'MRF'), with only virus occurrences included, 2) a Conditional Markov Random Field model (CRF) with only habitat- and host-related (collectively referred as 'environmental', see Table 1) variables included as additional constrains ('CRFenv'), 3) a CRF model with only host-related variables included as additional constraints ('CRFhost'), 4) a CRF model with only spatial variables and variables related to habitat (quality and weather) included as additional constraints ('CRFhabitat'), 5) a CRF model with only spatial variables included as additional constrains ('CRFspat'), 6) a Conditional Markov Random Field model with both all environmental as well as spatial variables included as additional constrains ('CRFfull'). We will refer to the variants (2-6) collectively as 'CRF models' or 'CRFs'.

We evaluated the model fit by calculating the Area Under Curve values[28] using the full data set. Following Clark *et al.*[25], we used cross validation (with four folds) to estimate model generality by comparing predicted and observed outcomes simultaneously for all species. To account for parameter uncertainty of the final model, we modelled 100 bootstrapped replicates for the model. If the 90% confidence interval of bootstrapped coefficients did not overlap with zero, we considered the variable to have a statistically significant effect. To test whether the associations between viruses are phylogenetically conservative, we compared the direct associations gained with all our network models to the phylogenetic relationships of the virus species, constructed from taxonomy, by conducting a Mantel test between the matrices.

All analyses, results and figures were produced with R (version 4.0.2[18]), and packages 'vegan'[29], 'MFRcov'[25], 'igraph'[30], along with their dependencies. An R package called 'meta17-network' including the analytical pipeline, data, and documentation for full reproduction of the results can be found in Github (aminorberg/meta17network-pkg).

References

1. Sagar, A. G. R. & Harper, J. L. Plantago Major L., P. Media L. and P. Lancoeolata L. *J. Ecol.* **52**, 189–221 (1964).
2. Ojanen, S. P., Nieminen, M., Meyke, E., Pöyry, J. & Hanski, I. Long-term metapopulation study of the Glanville fritillary butterfly (Melitaea cinxia): Survey methods, data management, and long-term population trends. *Ecol. Evol.* **3**, 3713–3737 (2013).
3. Hanski, I. *Metapopulation Ecology*. *Oxford Series in Ecology and Evolution* (Oxford University Press, Oxford, UK, 1999).
4. Chang, S., Puryear, J. & Cairney, J. A Simple and Efficient Method for Isolating RNA from Pine Trees. *Plant Mol. Biol. Report.* **11**, 113–116 (1993).
5. Cuellar, W. J. *et al.* Elimination of antiviral defense by viral RNase III. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 10354–10358 (2009).
6. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**, 1–18 (2018).
7. Zheng, Y. *et al.* VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **500**, 130–138 (2017).
8. Hulo, C. *et al.* ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**, 576–582 (2011).
9. Susi, H., Laine, A.-L., Filloux, D., Roumagnac, P. & Frilander, M. J. Diverse and variable virus communities in wild plant populations revealed by metagenomic tools. *PeerJ* **7**, e6140 (2019).
10. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).

11. Papaïx, J., Burdon, J. J., Zhan, J. & Thrall, P. H. Crop pathogen emergence and evolution in agro-ecological landscapes. *Evol. Appl.* **8**, 385–402 (2015).
12. Bernardo, P. *et al.* Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. *ISME J.* **12**, 173–184 (2018).
13. Susi, H. & Laine, A.-L. Agricultural land use disrupts biodiversity mediation of virus infections in wild plant populations. *New Phytol.* **230**, 2447- (2020).
14. QGIS Development Team. QGIS Geographic Information System. (2019).
15. Velásquez, A. C., Castroverde, C. D. M. & He, S. Y. Plant–Pathogen Warfare under Changing Climate Conditions. *Curr. Biol.* **28**, R619–R634 (2018).
16. Aalto, J., Pirinen, P. & Jylhä, K. New gridded daily climatology of Finland: permutation-based uncertainty estimates and temporal trends in climate. *J. Geophys. Res. Atmos.* **121**, (2016).
17. Dray, S., Legendre, P. & Peres-Neto, P. R. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Modell.* **196**, 483–493 (2006).
18. R Core Team. R: A language and environment for statistical computing. (2020).
19. Bivand, R. S., Pebesma, E. & Gomez-Rubio, V. *Applied spatial data analysis with R*. (Springer, NY, 2013).
20. Dray, S. *et al.* adespatial: Multivariate Multiscale Spatial Analysis. R package version 0.3-14. (2021).
21. Almeida-Neto, M., Guimarães, P., Guimarães, P. R., Loyola, R. D. & Ulrich, W. A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos* **117**, 1227–1239 (2008).
22. Stone, L. & Roberts, A. The Checkerboard Score and Species Distributions. *Oecologia* **85**, 74–79 (1990).
23. Scheiner, S. M. Six types of species-area curves. *Glob. Ecol. Biogeogr.* **12**, 441–447 (2003).
24. Hart, S. P., Usinowicz, J. & Levine, J. M. The spatial scales of species coexistence. *Nat. Ecol. Evol.* **1**, 1066–1073 (2017).
25. Clark, N. J., Wells, K. & Lindberg, O. Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology* **99**, 1277–1283 (2018).
26. Cheng, J., Levina, E., Wang, P. & Zhu, J. A sparse Ising model with covariates. *Biometrics* **70**, 943–953 (2014).
27. Susi, H. & Laine, A.-L. Agricultural land use disrupts biodiversity mediation of virus infections in wild plant populations. *New Phytol.* (2020) doi:10.1111/nph.17156.
28. Hanley, J. A. & Mcneil, B. J. The Meaning and Use of the Area under a Receiver Characteristic. *Radiology* **143**, 29–36 (1982).
29. Oksanen, J. *et al.* vegan: Community Ecology Package. R package version 2.5-6. (2019).
30. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Syst.* (2006).

Supplementary Results

Table S1. Mean predictive performance measures for all the models. The measures mentioned in the main text re highlighted with **bold** font.

Model	Quantiles based on bootstrapping	Mean proportion of true positives out of all predicted positives	Mean prop. of all true predictions out of the whole data	Mean prop. of true positives out of all real positives ('sensitivity')	Mean prop. of true negatives out of all real negatives ('specificity')
MRF	2.5%	0.15	0.73	0.33	0.76
	50%	0.21	**0.76**	0.47	0.80
	97.5%	0.27	0.80	0.59	0.84
CRFenv	2.5%	0.65	0.89	0.068	0.99
	50%	0.85	**0.91**	0.15	1.00
	97.5%	0.95	0.92	0.25	1.00
CRFhabitat	2.5%	0.70	0.90	0.11	0.99
	50%	0.85	**0.91**	0.20	1.00
	97.5%	0.95	0.93	0.31	1.00
CRFhost	2.5%	0.67	0.90	0.06	0.99
	50%	0.86	**0.91**	0.16	1.00
	97.5%	0.96	0.93	0.26	1.00
CRFspat	2.5%	0.71	0.90	0.09	0.99
	50%	0.87	**0.91**	0.2	1.00
	97.5%	0.96	0.93	0.31	1.00
CRFfull	2.5%	0.72	0.90	0.09	1.00
	50%	0.88	**0.91**	0.17	1.00
	97.5%	0.97	0.93	0.25	1.00

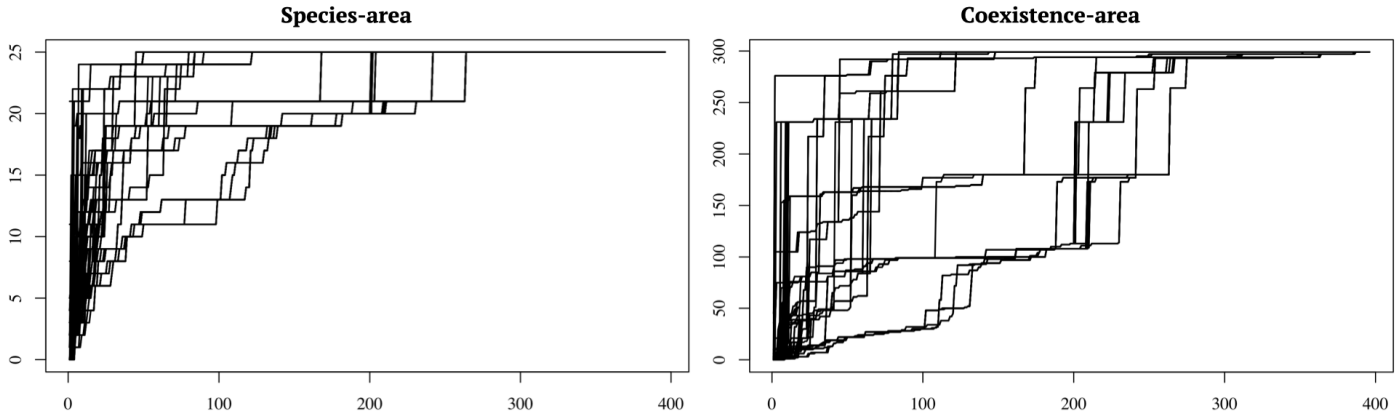

Figure S1. All 100 simulated species-area and species-coexistence curves. Each simulation (line) starts from a different host plant. After that, one plant is added to the pool at a time, always choosing the nearest one to the previous.

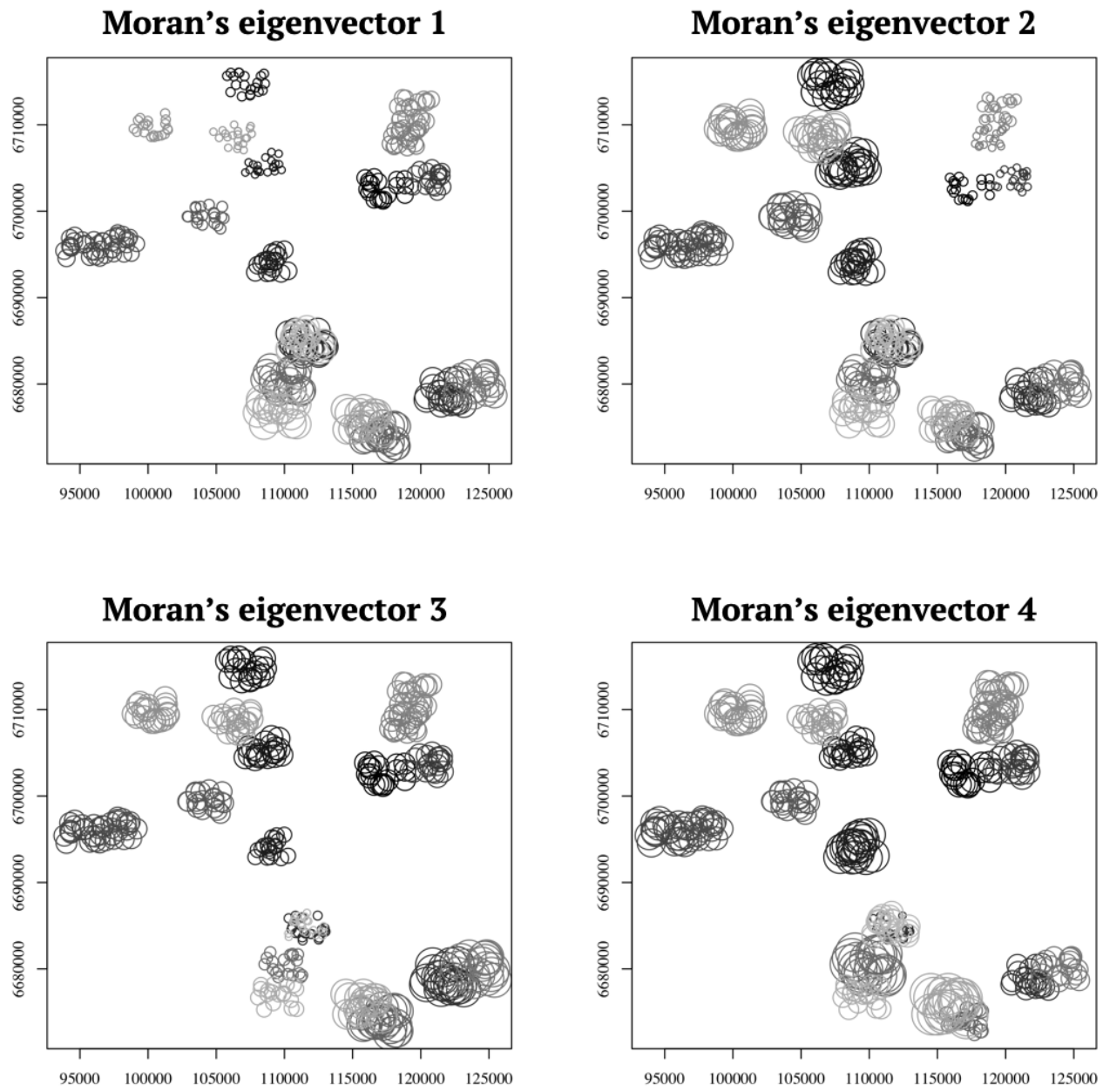

Figure S2. Moran's eigenvectors describing the spatial structure of the data. Each eigenvector is plotted in an individual panel. The horizontal axes show the X coordinates, vertical axes show the Y coordinates of the host plants. The circle size indicates the MEM value of that host plant.

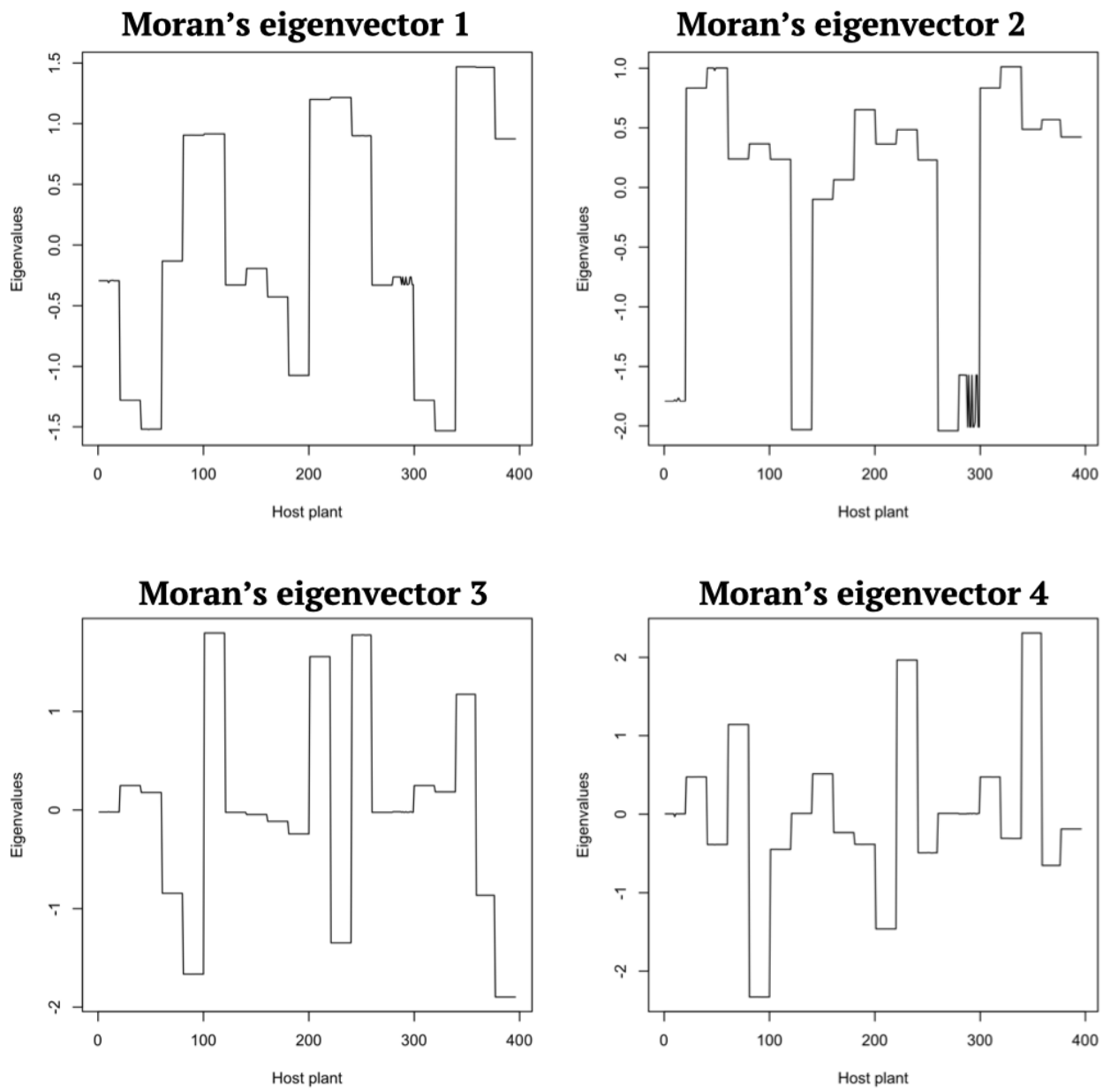

Figure S3. Moran's eigenvectors describing the spatial structure of the data. The horizontal axes show the MEM values.